DDS-3 AI and Machine Learning

Unit-I

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and act like humans. It involves the development of algorithms and computer programs that can perform tasks that typically require human intelligence such as visual perception, speech recognition, decision-making, and language translation. AI has the potential to revolutionize many industries and has a wide range of applications, from virtual personal assistants to self-driving cars.

Before leading to the meaning of artificial intelligence let understand what is the meaning of Intelligence-

Intelligence: The ability to learn and solve problems. This definition is taken from webster’s Dictionary.

The most common answer that one expects is “to make computers intelligent so that they can act intelligently!”, but the question is how much intelligent? How can one judge intelligence?

…as intelligent as humans. If the computers can, somehow, solve real-world problems, by improving on their own from past experiences, they would be called “intelligent”.

Thus, the AI systems are more generic(rather than specific), can “think” and are more flexible.

Intelligence, as we know, is the ability to acquire and apply knowledge. Knowledge is the information acquired through experience. Experience is the knowledge gained through exposure(training). Summing the terms up, we get artificial intelligence as the “copy of something natural(i.e., human beings) ‘WHO’ is capable of acquiring and applying the information it has gained through exposure.”

Intelligence is composed of:

• Reasoning

• Learning

• Problem-Solving

• Perception

• Linguistic Intelligence

Many tools are used in AI, including versions of search and mathematical optimization, logic, and methods based on probability and economics. The AI field draws upon computer science, mathematics, psychology, linguistics, philosophy, neuroscience, artificial psychology, and many others.

The main focus of artificial intelligence is towards understanding human behavior and performance. This can be done by creating computers with human-like intelligence and capabilities. This includes natural language processing, facial analysis and robotics. The main applications of AI are in military, healthcare, and computing; however, it’s expected that these applications will start soon and become part of our everyday lives.

Many theorists believe that computers will one day surpass human intelligence; they’ll be able to learn faster, process information more effectively and make decisions faster than humans. However, it’s still a work in progress as there are many limitations to how much artificial intelligence is achieved. For example, computers don’t perform well in dangerous or cold environments; they also struggle with physical tasks such as driving cars or operating heavy machinery. Even so, there are many exciting things ahead for artificial intelligence!

Uses of Artificial Intelligence :

Artificial Intelligence has many practical applications across various industries and domains, including:

1. Healthcare: AI is used for medical diagnosis, drug discovery, and predictive analysis of diseases.

2. Finance: AI helps in credit scoring, fraud detection, and financial forecasting.

3. Retail: AI is used for product recommendations, price optimization, and supply chain management.

4. Manufacturing: AI helps in quality control, predictive maintenance, and production optimization.

5. Transportation: AI is used for autonomous vehicles, traffic prediction, and route optimization.

6. Customer service: AI-powered chatbots are used for customer support, answering frequently asked questions, and handling simple requests.

7. Security: AI is used for facial recognition, intrusion detection, and cybersecurity threat analysis.

8. Marketing: AI is used for targeted advertising, customer segmentation, and sentiment analysis.

9. Education: AI is used for personalized learning, adaptive testing, and intelligent tutoring systems.

This is not an exhaustive list, and AI has many more potential applications in various domains and industries.

Need for Artificial Intelligence

1. To create expert systems that exhibit intelligent behavior with the capability to learn, demonstrate, explain, and advise its users.

2. Helping machines find solutions to complex problems like humans do and applying them as algorithms in a computer-friendly manner.

3. Improved efficiency: Artificial intelligence can automate tasks and processes that are time-consuming and require a lot of human effort. This can help improve efficiency and productivity, allowing humans to focus on more creative and high-level tasks.

4. Better decision-making: Artificial intelligence can analyze large amounts of data and provide insights that can aid in decision-making. This can be especially useful in domains like finance, healthcare, and logistics, where decisions can have significant impacts on outcomes.

5. Enhanced accuracy: Artificial intelligence algorithms can process data quickly and accurately, reducing the risk of errors that can occur in manual processes. This can improve the reliability and quality of results.

6. Personalization: Artificial intelligence can be used to personalize experiences for users, tailoring recommendations, and interactions based on individual preferences and behaviors. This can improve customer satisfaction and loyalty.

7. Exploration of new frontiers: Artificial intelligence can be used to explore new frontiers and discover new knowledge that is difficult or impossible for humans to access. This can lead to new breakthroughs in fields like astronomy, genetics, and drug discovery.

Approaches of AI

There are a total of four approaches of AI and that are as follows:

• Acting humanly (The Turing Test approach): This approach was designed by Alan Turing. The ideology behind this approach is that a computer passes the test if a human interrogator, after asking some written questions, cannot identify whether the written responses come from a human or from a computer.

• Thinking humanly (The cognitive modeling approach): The idea behind this approach is to determine whether the computer thinks like a human.

• Thinking rationally (The “laws of thought” approach): The idea behind this approach is to determine whether the computer thinks rationally i.e. with logical reasoning.

• Acting rationally (The rational agent approach): The idea behind this approach is to determine whether the computer acts rationally i.e. with logical reasoning.

• Machine Learning approach: This approach involves training machines to learn from data and improve performance on specific tasks over time. It is widely used in areas such as image and speech recognition, natural language processing, and recommender systems.

• Evolutionary approach: This approach is inspired by the process of natural selection in biology. It involves generating and testing a large number of variations of a solution to a problem, and then selecting and combining the most successful variations to create a new generation of solutions.

• Neural Networks approach: This approach involves building artificial neural networks that are modeled after the structure and function of the human brain. Neural networks can be used for tasks such as pattern recognition, prediction, and decision-making.

• Fuzzy logic approach: This approach involves reasoning with uncertain and imprecise information, which is common in real-world situations. Fuzzy logic can be used to model and control complex systems in areas such as robotics, automotive control, and industrial automation.

• Hybrid approach: This approach combines multiple AI techniques to solve complex problems. For example, a hybrid approach might use machine learning to analyze data and identify patterns, and then use logical reasoning to make decisions based on those patterns.

Applications of AI include Natural Language Processing, Gaming, Speech Recognition, Vision Systems, Healthcare, Automotive, etc.

Forms of AI:

1) Weak AI:

• Weak AI is an AI that is created to solve a particular problem or perform a specific task.

• It is not a general AI and is only used for specific purpose.

• For example, the AI that was used to beat the chess grandmaster is a weak AI as that serves only 1 purpose but it can do it efficiently.

2) Strong AI:

• Strong AI is difficult to create than weak AI.

• It is a general purpose intelligence that can demonstrate human abilities.

• Human abilities such as learning from experience, reasoning, etc. can be demonstrated by this AI.

3) Super Intelligence

• As stated by a leading AI thinker Nick Bostrom, “Super Intelligence is an AI that is much smarter than the best human brains in practically every field”.

• It ranges from a machine being just smarter than a human to a machine being trillion times smarter than a human.

• Super Intelligence is the ultimate power of AI.

An AI system is composed of an agent and its environment. An agent(e.g., human or robot) is anything that can perceive its environment through sensors and acts upon that environment through effectors. Intelligent agents must be able to set goals and achieve them. In classical planning problems, the agent can assume that it is the only system acting in the world, allowing the agent to be certain of the consequences of its actions. However, if the agent is not the only actor, then it requires that the agent can reason under uncertainty. This calls for an agent that cannot only assess its environment and make predictions but also evaluate its predictions and adapt based on its assessment. Natural language processing gives machines the ability to read and understand human language. Some straightforward applications of natural language processing include information retrieval, text mining, question answering, and machine translation. Machine perception is the ability to use input from sensors (such as cameras, microphones, sensors, etc.) to deduce aspects of the world. e.g., Computer Vision. Concepts such as game theory, and decision theory, necessitate that an agent can detect and model human emotions.

Many times, students get confused between Machine Learning and Artificial Intelligence, but Machine learning, a fundamental concept of AI research since the field’s inception, is the study of computer algorithms that improve automatically through experience. The mathematical analysis of machine learning algorithms and their performance is a branch of theoretical computer science known as a computational learning theory.

Stuart Shapiro divides AI research into three approaches, which he calls computational psychology, computational philosophy, and computer science. Computational psychology is used to make computer programs that mimic human behavior. Computational philosophy is used to develop an adaptive, free-flowing computer mind. Implementing computer science serves the goal of creating computers that can perform tasks that only people could previously accomplish.

AI has developed a large number of tools to solve the most difficult problems in computer science, like:

• Search and optimization

• Logic

• Probabilistic methods for uncertain reasoning

• Classifiers and statistical learning methods

• Neural networks

• Control theory

• Languages

High-profile examples of AI include autonomous vehicles (such as drones and self-driving cars), medical diagnosis, creating art (such as poetry), proving mathematical theorems, playing games (such as Chess or Go), search engines (such as Google search), virtual assistants (such as Siri), image recognition in photographs, spam filtering, prediction of judicial decisions[204] and targeted online advertisements. Other applications include Healthcare, Automotive, Finance, Video games, etc

Are there limits to how intelligent machines – or human-machine hybrids – can be? A superintelligence, hyperintelligence, or superhuman intelligence is a hypothetical agent that would possess intelligence far surpassing that of the brightest and most gifted human mind. ‘‘Superintelligence’’ may also refer to the form or degree of intelligence possessed by such an agent.

This article is contributed by Palak Jain. If you like GeeksforGeeks and would like to contribute, you can also write an article using write.geeksforgeeks.org or mail your article to review-team@geeksforgeeks.org. See your article appearing on the GeeksforGeeks main page and help other Geeks.

Please write comments if you find anything incorrect, or if you want to share more information about the topic discussed above.

Drawbacks of Artificial Intelligence :

1. Bias and unfairness: AI systems can perpetuate and amplify existing biases in data and decision-making.

2. Lack of transparency and accountability: Complex AI systems can be difficult to understand and interpret, making it challenging to determine how decisions are being made.

3. Job displacement: AI has the potential to automate many jobs, leading to job loss and a need for reskilling.

4. Security and privacy risks: AI systems can be vulnerable to hacking and other security threats, and may also pose privacy risks by collecting and using personal data.

5. Ethical concerns: AI raises important ethical questions about the use of technology for decision-making, including issues related to autonomy, accountability, and human dignity.

Technologies Based on Artificial Intelligence:

1. Machine Learning: A subfield of AI that uses algorithms to enable systems to learn from data and make predictions or decisions without being explicitly programmed.

2. Natural Language Processing (NLP): A branch of AI that focuses on enabling computers to understand, interpret, and generate human language.

3. Computer Vision: A field of AI that deals with the processing and analysis of visual information using computer algorithms.

4. Robotics: AI-powered robots and automation systems that can perform tasks in manufacturing, healthcare, retail, and other industries.

5. Neural Networks: A type of machine learning algorithm modeled after the structure and function of the human brain.

6. Expert Systems: AI systems that mimic the decision-making ability of a human expert in a specific field.

7. Chatbots: AI-powered virtual assistants that can interact with users through text-based or voice-based interfaces.

Applications

Issues of Artificial Intelligence :

Artificial Intelligence has the potential to bring many benefits to society, but it also raises some important issues that need to be addressed, including:

1. Bias and Discrimination: AI systems can perpetuate and amplify human biases, leading to discriminatory outcomes.

2. Job Displacement: AI may automate jobs, leading to job loss and unemployment.

3. Lack of Transparency: AI systems can be difficult to understand and interpret, making it challenging to identify and address bias and errors.

4. Privacy Concerns: AI can collect and process vast amounts of personal data, leading to privacy concerns and the potential for abuse.

5. Security Risks: AI systems can be vulnerable to cyber attacks, making it important to ensure the security of AI systems.

6. Ethical Considerations: AI raises important ethical questions, such as the acceptable use of autonomous weapons, the right to autonomous decision making, and the responsibility of AI systems for their actions.

7. Regulation: There is a need for clear and effective regulation to ensure the responsible development and deployment of AI.

It’s crucial to address these issues as AI continues to play an increasingly important role in our lives and society.

The Future of AI Technologies:

1. Reinforcement Learning: Reinforcement Learning is an interesting field of Artificial Intelligence that focuses on training agents to make intelligent decisions by interacting with their environment.

2. Explainable AI: this AI techniques focus on providing insights into how AI models arrive at their conclusions.

3. Generative AI: Through this technique AI models can learn the underlying patterns and create realistic and novel outputs.

4. Edge AI:AI involves running AI algorithms directly on edge devices, such as smartphones, IoT devices, and autonomous vehicles, rather than relying on cloud-based processing.

5. Quantum AI: Quantum AI combines the power of quantum computing with AI algorithms to tackle complex problems that are beyond the capabilities of classical computers.

Essential concept in artificial intelligence:

## **What are the Fundamental AI Concepts?**

To fully understand how AI works, you need to learn about the following basic concepts first:

Machine Learning

**Machine Learning (ML)** is a branch of Artificial Intelligence that is based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

In his book *The Master Algorithm*, world expert and leading researcher Pedro Domingos describes 5 tribes or currents of Machine Learning, that he divides based on their interests or origins:

* **Symbolists:** They are based on logic and philosophy, and practice inverse deduction
* **Connectionists:** They rely on neuroscience and try to connect small brains from what they call*back programming*to create a neural network that can interpret the data from their interconnections.  Deep Learning comes from here.
* **Evolutionaries:** They are based on evolutionary biology and try to apply genome and DNA evolution principles, claiming that algorithms will evolve and adapt to unknown conditions and processes.
* **Analogizers:** They rely on psychology and see analogy as a basis for solving any problems that may arise.
* **Bayesians:** They are based on statistics and probability. His algorithm, probabilistic inference, learns by trying to calculate how improbable a fact is to rule it out as a possible solution.

We also find 4 types of Machine Learning depending on the need for human supervision:

* **Supervised Learning:** Itlearns by receiving a lot of labeled training data that allows generalizing in new cases.
* **Unsupervised Learning:** It learns by observing, understanding, and abstracting patterns directly from the information. It is very similar to how we humans think.
* **Semi-supervised Learning:** It learns based on both labeled and unlabeled training data, with the proportion of unlabeled data typically being larger.
* **Reinforcement Learning:**Itlearns from experience through trial-error and reward-punishment. This technique is being widely studied since it does not require large amounts of data.

### ****1. Machine Learning****

In the simplest terms, machine learning (ML) is a subset of AI. Its core lies in the idea that computer systems can learn on their own from data obtained from performing previous tasks and past experiences. That means that you don’t have to pre-program an AI device every time you need it to work on a job.

ML has three subcategories—supervised, unsupervised, and reinforcement.

* [Supervised learning](https://www.techslang.com/definition/what-is-supervised-learning/) occurs when an AI system arrives at a predictable conclusion based on existing data.
* [Unsupervised learning](https://www.techslang.com/definition/what-is-unsupervised-learning/), meanwhile, takes place when the AI agent produces an unpredictable result, which it was not pre-trained to do.
* [Reinforcement learning](https://www.techslang.com/definition/what-is-reinforcement-learning/) (also known as “goal-oriented programming”) deals with training the AI algorithm to recognize rewards and punishments so that it can come up with the best solution to a problem.

### ****2. Deep Learning****

This takes ML up a notch. This subset of AI refers to a system’s ability to take unstructured data from multiple sources, analyze it, and apply it to solve new problems. Deep learning is also known as “differential programming.”

### ****3. Artificial neural network (ANN)****

[Artificial neural network](https://www.techslang.com/how-does-an-artificial-neural-network-work/) refers to a system or an algorithm used in deep learning that mimics how the human brain’s neural circuits function, such as when making sense of things and events.

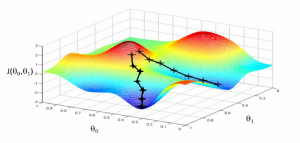
## **What are Other Relevant Concepts in AI?**

Listed below are more AI-related terms that can deepen your understanding.

* **Categorization:** Building a successful AI system requires creating a type of category or benchmark for a specific field. These criteria or metrics are used by the machine to diagnose a problem. After further analysis, its diagnosis could eventually lead to a fitting solution.
* **Classification:** This is a property of an AI model that points to its ability to “classify” what type of problem it encounters, what causes it, and what solution can best remedy it. In medical diagnostics, for example, an AI-powered tool identifies an illness based on its unique qualities.
* **Collaborative filtering:** This refers to the capability of an AI system to make decisions or give recommendations on its own based on what it learned from a user’s past preferences and actions. An example of its results is a recommendation that you receive via ads or media platforms.
* [**Natural language processing (NLP)**](https://www.techslang.com/definition/what-is-natural-language-processing-nlp/)**:** This is a characteristic of advanced neural networks that describes their ability to interpret tasks and produce outputs that humans can read. The term also pertains to a field of computer science that focuses on developing computers that can understand natural language through reading and listening. Conversational AI platforms, such as messaging apps and chatbots, use NLP.
* [**Data mining**](https://www.techslang.com/definition/what-is-data-mining/)**:** This involves extracting unstructured data from various databases and websites to enrich predictive AI algorithms. AI systems use statistical methods to analyze aggregated data for trends and associations, which, in turn, allow them to generate new information.

# Most Common Machine Learning Tasks

[December 4, 2022](https://vitalflux.com/7-common-machine-learning-tasks-related-methods/) by [Ajitesh Kumar](https://vitalflux.com/author/vitalflux/) · [Leave a comment](https://vitalflux.com/7-common-machine-learning-tasks-related-methods/#respond)



This article represents some of the **most common machine learning tasks** that one may come across while trying to solve **machine learning** problems. Also listed is a set of [**machine learning**](https://vitalflux.com/category/machine-learning/)**methods** that could be used to resolve these tasks. Please feel free to comment/suggest if I missed mentioning one or more important points. Also, sorry for the typos.

You might want to check out the post on [what is machine learning?](https://vitalflux.com/what-is-machine-learning-concepts-examples/). Different aspects of machine learning concepts have been explained with the help of examples. Here is an excerpt from the page:

*Machine learning is about approximating mathematical functions (equations) representing real-world scenarios. These mathematical functions are also referred to as “mathematical models” or just models.*

Following are the key machine learning tasks briefed later in this article:

1. Regression
2. Classification
3. Clustering
4. Transcription
5. Machine translation
6. Anomaly detection
7. Synthesis & sampling
8. Estimation of probability density and probability mass function
9. Similarity matching
10. Co-occurrence grouping
11. Causal modeling
12. Link profiling

Following are the key development phases that are used to solve the different tasks listed above. These form the key phases of the machine learning models’ (MLM) development lifecycle.

1. Data gathering
2. Data preprocessing
3. Exploratory data analysis (EDA)
4. Feature engineering including feature creation/extraction, feature selection, dimensionality reduction
5. Training machine learning models
6. Model / Algorithm selection
7. Testing and matching
8. Model monitoring
9. Model retraining

Table of Contents

* [Machine Learning Tasks](https://vitalflux.com/7-common-machine-learning-tasks-related-methods/#Machine_Learning_Tasks)
* [Development phases of MLM Development Lifecycle](https://vitalflux.com/7-common-machine-learning-tasks-related-methods/#Development_phases_of_MLM_Development_Lifecycle)

## **Machine Learning Tasks**

The following are some of the key tasks which can be performed using machine learning models:

* **Regression**: Regression tasks mainly deal with the estimation of numerical values (**continuous variables**). Regression task in machine learning is a supervised machine learning technique used to predict the values of a given target variable based on the input of one or more independent variables. It is a task of fitting a mathematical model to observed data points, where the objective is to minimize the sum of squared errors between the observed data and the predicted values. In regression tasks, we use linear and non-linear models to build our predictive models. Linear models have a basic assumption that there exists a linear relationship between the input and output variables, while non-linear models do not rely on any such assumptions. The goal of linear regression is to find the best fit line for our data whereas in non-linear regression, we try to identify complex relationships within our dataset. In addition to prediction problems, regression tasks can also be used for inference problems where we want to understand how variables interact with each other. For example, you might use a linear regression model in order to investigate how two or more independent variables influence a dependent variable such as salary or job satisfaction. This allows us to gain insights about how changes in one independent variable affects another independent variable or even the dependent variable itself without having to make any changes directly ourselves.  
  Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:
  + - Kernel regression (Higher accuracy)
    - Gaussian process regression (Higher accuracy)
    - Regression trees
    - Linear regression
    - Support vector regression
    - LASSO / Ridge
    - Deep learning
    - Random forests
* **Classification**: Classification tasks are simply related to predicting a category of data (**discrete variables**). Classification is a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics. It’s one of the most common types of supervised learning techniques and it’s used for everything from banking fraud detection to face recognition. At its core, classification is the process of sorting items into two or more mutually exclusive groups, often called classes. The goal is to correctly assign a given item to the right class based on its features. So in order to do this, the machine must learn how to differentiate between different classes using these features. This can be done by pre-labeling data so that the computer knows which class each item belongs to, then training it with different algorithms until it learns how to accurately identify each class. Simply speaking, a classification task results in the model which, given a new individual, determines which class that individual belongs to. A closely related task is **scoring** or class **probability estimation**. A scoring model applied to an individual produces, instead of a class prediction, a score representing the probability (or some other quantification of likelihood) that that individual belongs to each class. One of the most common examples is predicting whether or not an email if spam or ham. Some of the common use cases could be found in the area of healthcare such as whether a person is suffering from a particular disease or not. It also has its application in financial use cases such as determining whether a transaction is a fraud or not. You might want to check this page on real-world examples of classification models, [machine learning classification models real-life examples](https://vitalflux.com/classification-problems-real-world-examples/). The ML methods such as the following could be applied to solve classification tasks:
  + Kernel discriminant analysis (Higher accuracy)
  + K-Nearest Neighbors (Higher accuracy)
  + Artificial neural networks (ANN) (Higher accuracy)
  + Support vector machine (SVM) (Higher accuracy)
  + Random forests (Higher accuracy)
  + Decision trees
  + Boosted trees
  + Logistic regression
  + naive Bayes
  + Deep learning
* **Clustering**: Clustering is a commonly used machine learning task in which data points are grouped into clusters, or groups of closely related data points. It is an unsupervised approach that does not require labeled data and can be used to identify patterns or similarities within a dataset. Clustering has many applications ranging from customer segmentation, market segmentation, image segmentation, document classification, and more. At its core, clustering is a process of partitioning a set of objects into distinct groups such that the elements in each group are similar to each other while those belonging to different groups are very dissimilar. The following are four different type of clustering algorithms:
  + Prototype based clustering (K-means)
  + Hierarchical clustering
  + DBSCAN (Density based spatial clustering of applications with noise)
  + Distribution based clustering
* **Similarity matching**: Similarity matching task in machine learning is a task in which machines are trained to match items based on their similarity. This type of task can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems. In order for a machine learning system to be able to perform well at this type of task, it needs to be able to learn how to distinguish between similar and dissimilar items. This can be done by creating feature vectors from examples of known data points, then using that information as training data so that the machine can make accurate predictions when presented with new data points. Once trained, similarity matching algorithms can be used in many ways such as providing recommendations or helping with search engine optimization. For example, if you were looking for a particular product online but couldn’t find it through traditional search methods, **similarity matching could help by presenting other products that closely match your desired item** based on their features and characteristics. Similarly, recommendation systems use such **algorithms to suggest items that users may like based on their past preferences** or those of other users with similar interests.
* **Co-occurrence grouping**: Co-occurrence grouping tasks are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks. In this task, association between entities are found based on transactions involving them. for example, w*hat items are commonly purchased together*? The difference between clustering and co-occurrence grouping is that in clustering, the similarity between objects is found based on the objects’ attributes while in co-occurrence grouping, the similarity of objects is found based on them appearing together in transactions. For example, the purchase records from a supermarket may uncover the association that bread is purchased together with eggs much more frequently than expected.
* **Multivariate querying**: Multivariate querying is about querying or finding similar objects. Some of the following ML methods could be used for such problems:
  + Nearest neighbors
  + Range search
  + Farthest neighbors
* **Probability density and mass function estimation**: Probability density function estimation problems are related to finding the likelihood or frequency of objects. In probability and statistics, density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function. Some of the following ML methods could be used for solving density estimation tasks:
  + Kernel density estimation (Higher accuracy)
  + Mixture of Gaussians
  + Density estimation tree
* **Machine translation**: Machine translation is the process of translating text from one language to another using machine learning algorithms. There are many different machine translation tasks, such as machine translation of documents, machine translation of the speech, and machine translation of web pages. Deep learning models have achieved state-of-the-art results on many machine translation tasks. For example, deep learning models have been used to machine translate Web pages from English to Chinese with close to human-level accuracy. In addition, deep learning models have been used to machine translate speech from English to French with close to human-level accuracy. Machine translation is an important application of machine learning that has the potential to change the way people communicate with each other.
* **Anomaly detection**: Anomaly detection is the process of identifying unusual patterns in data that do not conform to expected behavior. It is often used in a wide range of applications, such as detecting fraudulent activity in financial data, detecting malicious behavior in network traffic data, and identifying equipment malfunctions in sensor data. Anomaly detection can be performed using a variety of machine learning models, such as density-based methods, cluster-based methods, and rule-based methods. Each of these methods has its own strengths and weaknesses, so it is important to select the right model for the particular application. Anomaly detection using machine learning is a complex process, but it can be extremely effective at identifying rare events that would otherwise be overlooked.
* **Synthesis & sampling**: Synthesis and sampling are essential tasks in deep learning and machine learning. They are used to generate new data from existing data or to select a representative subset of data for further analysis. Synthesis and sampling are often used together, in order to create a more diverse and representative dataset. Synthesis can be used to generate new data points, by extrapolating from existing data points. For example, if we have a dataset of images of animals, we can use synthesis to generate new images of animals that are similar to the ones in the dataset. Sampling can be used to select a subset of data that is representative of the entire dataset. For example, if we have a dataset of images of animals, we can use sampling to select a subset of images that represents all the different animal types in the dataset.
* **Transcription**: Transcription tasks are those that involve converting audio or video recordings or images having text into written text. They are commonly used in fields such as journalism, academia, and medicine. In recent years, transcription tasks have been automated to some degree using machine learning algorithms. Deep learning models can be trained to transcribe audio recordings with a high degree of accuracy. However, these models require a large amount of data to train on, and they often struggle with background noise and accents. As a result, transcription is still largely a manual task. Professional transcriptionists use their knowledge of the language and their attention to detail to produce accurate transcription services.
* **Causal modeling**: Causal modeling is a type of machine learning task that aims to infer the causes and effects of certain conditions or variables. It is an important tool for researchers in fields such as epidemiology, economics, psychology, marketing, and political science. In causal modeling, data is used to make inferences about the relationships between variables. The goal is to identify which variables are causing certain outcomes and how they are related. For example, a researcher may want to determine the causal relationship between smoking cigarettes and lung cancer. Does smoking cigarettes is related to lung cancer?
* **Dimensionality reduction (feature extraction)**: As per the [Wikipedia page on Dimension reduction](http://en.wikipedia.org/wiki/Dimensionality_reduction), Dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. Following are some ML methods that could be used for dimension reduction:
  + Manifold learning/KPCA (Higher accuracy)
  + Principal component analysis
  + Independent component analysis
  + Gaussian graphical models
  + Non-negative matrix factorization
  + Compressed sensing
* **Link prediction**: Link prediction task in machine learning is a task that focuses on identifying potential connections between entities that are not yet connected. It is used to predict relationships between entities, such as customers, products, authors, and more. The goal of link prediction is to build a model that can accurately identify connections between entities in a dataset. In the past, link prediction has primarily been used for social network analysis where it’s often used to suggest friends or followers for users of a social network platform. However, it can also be applied to other types of data such as customer transactions data or scientific research papers. Link prediction models are also often used in recommendation systems to recommend items to customers based on their past behavior or preferences. **Link prediction can also estimate the strength of a link.** For example, for recommending movies to customers, link prediction can be used to create a graph between customers and the movies they’ve watched or rated. Within the graph, those potential links (strong link) are searched that should exist between customers and movies.

## **Development phases of MLM Development Lifecycle**

Following are the most common machine learning models development phases that one could come across most frequently while solving different machine learning tasks:

1. **Data Gathering**: Any machine learning problem requires a lot of data for training/testing purposes. Identifying the right data sources and gathering data from these data sources is the key. Data could be found from databases, external agencies, the internet, etc.
2. **Data Preprocessing**: Before starting training the models, it is of utmost importance to prepare data appropriately. As part of data preprocessing, some of the following is done:
   * **Data cleaning**: Data cleaning requires one to identify attributes having not enough data or attributes which are not have variance. These data (rows and columns) need to be removed from the training data set.
   * **Missing data imputation**: Handling missing data using data imputation techniques such as replacing missing data with mean, median, or mode. Here is my post on this topic: [Replace missing values with mean, median or mode](https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/)
3. **Exploratory Data Analysis (EDA)**: Once data is preprocessed, the next step is to perform exploratory data analysis to understand data distribution and relationships between/within the data. Some of the following are performed as part of EDA:
   * Correlation analysis
   * Multicollinearity analysis
   * Data distribution analysis
4. **Feature Engineering**: Feature engineering is one of the critical tasks which would be used when building machine learning models. Feature engineering is important because selecting the right features would not only help build models of higher accuracy but also help achieve objectives related to building simpler models, reducing overfitting, etc. Feature engineering includes tasks such as deriving features from raw features, identifying important features, feature extraction, and feature selection. The following are some of the techniques which could be used for feature selection:
   * Filter methods help in selecting features based on the outcomes of statistical tests. The following are some of the statistical tests which are used:
     + Pearson’s correlation
     + Linear discriminant analysis (LDA)
     + Analysis of Variance (ANOVA)
     + Chi-square tests
   * Wrapper methods help in feature selection by using a subset of features and determining the model accuracy. The following are some of the algorithms used:
     + Forward selection
     + Backward elimination
     + Recursive feature elimination
   * Regularization techniques penalize one or more features appropriately to come up with most important features. The following are some of the algorithms used:
     + LASSO (L1) regularization
     + Ridge (L2) regularization
     + Elastic net regularization
     + Regularization with classification algorithms such as Logistic regression, SVM, etc.
5. **Training Models:**Once some of the features are determined, then comes training models with data related to those features. One popular algorithm used for while training regression models is **Gradient Descent** which helps us to find optimal parameter values in order to minimize our cost function (also known as error rate). In this method, we start with random initial parameter values and then gradually update them by taking small steps until we reach an optimal solution. This iterative process helps us reduce error rates over time and ultimately provide better predictions for our target variable.
6. **Model selection / Algorithm selection**: Many times, there are multiple models which are trained using different algorithms. One of the important tasks is to select the most optimal models for deploying them in production. **Hyperparameter tuning** is the most common task performed as part of model selection. Also, if there are two models trained using different algorithms which have similar performance, then one also needs to perform algorithm selection.
7. **Testing and matching**: Testing and matching tasks relate to comparing data sets. Following are some of the methods that could be used for such kinds of problems:
   * Minimum spanning tree
   * Bipartite cross-matching
   * N-point correlation
8. **Model monitoring**: Once the models are trained and deployed, they require to be monitored at regular intervals. Monitoring models require the processing actual values and predicted values and measuring the model performance based on appropriate metrics.
9. **Model retraining**: In case, the model performance degrades, the models are required to be retrained.  The following gets done as part of model retraining:
   * New features get determined
   * New algorithms can be used
   * Hyperparameters can get tuned
   * Model ensembles may get deployed

Unit-II

Supervised Learning:

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

## **How Supervised Learning Works?**

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

The working of Supervised learning can be easily understood by the below example and diagram:



Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

* If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square**.
* If the given shape has three sides, then it will be labelled as a **triangle**.
* If the given shape has six equal sides then it will be labelled as **hexagon**.

Now, after training, we test our model using the test set, and the task of the model is to identify the shape.

The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

## **Steps Involved in Supervised Learning:**

* First Determine the type of training dataset
* Collect/Gather the labelled training data.
* Split the training dataset into training **dataset, test dataset, and validation dataset**.
* Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
* Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
* Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
* Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

## **Types of supervised Machine learning Algorithms:**

Supervised learning can be further divided into two types of problems:



**1. Regression**

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

* Linear Regression
* Regression Trees
* Non-Linear Regression
* Bayesian Linear Regression
* Polynomial Regression

**2. Classification**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

* Random Forest
* Decision Trees
* Logistic Regression
* Support vector Machines

#### **Note: We will discuss these algorithms in detail in later chapters.**

## **Advantages of Supervised learning:**

* With the help of supervised learning, the model can predict the output on the basis of prior experiences.
* In supervised learning, we can have an exact idea about the classes of objects.
* Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

## **Disadvantages of supervised learning:**

* Supervised learning models are not suitable for handling the complex tasks.
* Supervised learning cannot predict the correct output if the test data is different from the training dataset.
* Training required lots of computation times.
* In supervised learning, we need enough knowledge about the classes of object.

# K-Nearest Neighbor(KNN) Algorithm

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
* **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



## **Why do we need a K-NN Algorithm?**

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



## **How does K-NN work?**

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



* Firstly, we will choose the number of neighbors, so we will choose the k=5.
* Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



* By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



* As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

## **How to select the value of K in the K-NN Algorithm?**

Below are some points to remember while selecting the value of K in the K-NN algorithm:

* There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
* A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
* Large values for K are good, but it may find some difficulties.

## **Advantages of KNN Algorithm:**

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

## **Disadvantages of KNN Algorithm:**

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data points for all the training samples.
* K-NN is computationally expensive.
* It is a lazy learner i.e. it uses all the training data at the runtime and hence is slow.
* Complexity is O(n) for each instance to be classified.
* Always needs to determine the value of K which may be complex some time.
* The algorithm gets significantly slower as the number of examples and predictors variables increase.

Parsing and importing data from a text file, Creating scatter plots with Matplotlib, normalizing numeric values

# Decision Tree

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.

During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split.

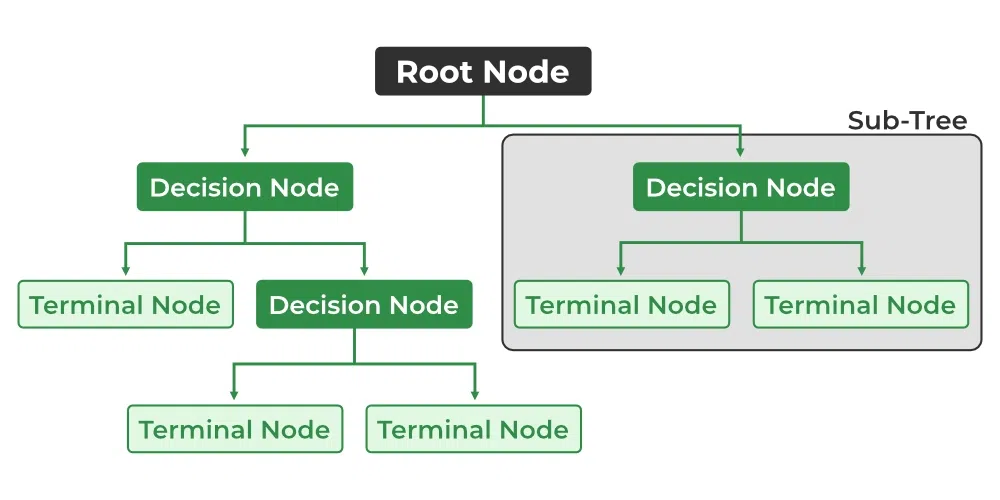
## What is a Decision Tree?

A decision tree is a flowchart-like [tree structure](https://www.geeksforgeeks.org/introduction-to-tree-data-structure-and-algorithm-tutorials/) where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile [supervised machine-learning](https://www.geeksforgeeks.org/ml-types-learning-supervised-learning/) algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. And it is also used in Random Forest to train on different subsets of training data, which makes random forest one of the most powerful algorithms in [machine learning](https://www.geeksforgeeks.org/machine-learning/).

### Decision Tree Terminologies

Some of the common Terminologies used in Decision Trees are as follows:

* **Root Node:** It is the topmost node in the tree,  which represents the complete dataset. It is the starting point of the decision-making process.
* Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
* **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
* **Splitting:**The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
* **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
* **Parent Node:** The node that divides into one or more child nodes.
* **Child Node:**The nodes that emerge when a parent node is split.
* **Impurity**: A measurement of the target variable’s homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The **Gini index** and **entropy** are two commonly used impurity measurements in decision trees for classifications task
* **Variance**: Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees. **Mean squared error, Mean Absolute Error, friedman\_mse, or Half Poisson deviance** are used to measure the variance for the regression tasks in the decision tree.
* **Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets
* **Pruning**: The process of removing branches from the tree that do not provide any additional information or lead to overfitting.



### Attribute Selection Measures:

**Construction of Decision Tree:** A tree can be “learned” by splitting the source set into subsets based on Attribute Selection Measures. Attribute selection measure (ASM) is a criterion used in decision tree algorithms to evaluate the usefulness of different attributes for splitting a dataset. The goal of ASM is to identify the attribute that will create the most homogeneous subsets of data after the split, thereby maximizing the information gain. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data.

#### Entropy:

Entropy is the measure of the degree of randomness or uncertainty in the dataset. In the case of classifications, It measures the randomness based on the distribution of class labels in the dataset.

The entropy for a subset of the original dataset having K number of classes for the ith node can be defined as:

Where,

* S is the dataset sample.
* k is the particular class from K classes
* p(k) is the proportion of the data points that belong to class k to the total number of data points in dataset sample S.
* Here p(i,k) should not be equal to zero.

**Important points related to Entropy:**

1. The entropy is 0 when the dataset is completely homogeneous, meaning that each instance belongs to the same class. It is the lowest entropy indicating no uncertainty in the dataset sample.
2. when the dataset is equally divided between multiple classes, the entropy is at its maximum value. Therefore, entropy is highest when the distribution of class labels is even, indicating maximum uncertainty in the dataset sample.
3. Entropy is used to evaluate the quality of a split. The goal of entropy is to select the attribute that minimizes the entropy of the resulting subsets, by splitting the dataset into more homogeneous subsets with respect to the class labels.
4. The highest information gain attribute is chosen as the splitting criterion (i.e., the reduction in entropy after splitting on that attribute), and the process is repeated recursively to build the decision tree.

#### Gini Impurity or index:

Gini Impurity is a score that evaluates how accurate a split is among the classified groups. The Gini Impurity evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes. In this case, we want to have a Gini index score as low as possible. Gini Index is the evaluation metric we shall use to evaluate our Decision Tree Model.

Here,

* pi is the proportion of elements in the set that belongs to the ith category.

#### Information Gain:

Information gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable. The higher the information gain, the more valuable the feature is in predicting the target variable.

The information gain of an attribute A, with respect to a dataset S, is calculated as follows:

where

* A is the specific attribute or class label
* |H| is the entropy of dataset sample S
* |HV| is the number of instances in the subset S that have the value v for attribute A

Information gain measures the reduction in entropy or variance achieved by partitioning the dataset on attribute A. The attribute that maximizes information gain is chosen as the splitting criterion for building the decision tree.

Information gain is used in both classification and regression decision trees. In classification, entropy is used as a measure of impurity, while in regression, variance is used as a measure of impurity. The information gain calculation remains the same in both cases, except that entropy or variance is used instead of entropy in the formula.

**How does the Decision Tree algorithm Work?**  
The decision tree operates by analyzing the data set to predict its classification. It commences from the tree’s root node, where the algorithm views the value of the root attribute compared to the attribute of the record in the actual data set. Based on the comparison, it proceeds to follow the branch and move to the next node.

The algorithm repeats this action for every subsequent node by comparing its attribute values with those of the sub-nodes and continuing the process further. It repeats until it reaches the leaf node of the tree. The complete mechanism can be better explained through the algorithm given below.

* Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
* Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
* Step-3: Divide the S into subsets that contains possible values for the best attributes.
* Step-4: Generate the decision tree node, which contains the best attribute.
* Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf nodeClassification and Regression Tree algorithm.

**Advantages of the Decision Tree:**

1. It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
2. It can be very useful for solving decision-related problems.
3. It helps to think about all the possible outcomes for a problem.
4. There is less requirement of data cleaning compared to other algorithms.

**Disadvantages of the Decision Tree:**

1. The decision tree contains lots of layers, which makes it complex.
2. It may have an overfitting issue, which can be resolved using the Random Forest algorithm.
3. For more class labels, the computational complexity of the decision tree may increase.

**What are appropriate problems for Decision tree learning?**

Although a variety of decision tree learning methods have been developed with somewhat differing capabilities and requirements, decision tree learning is generally best suited to problems with the following characteristics:

**1. Instances are represented by attribute-value pairs:**

In the world of decision tree learning, we commonly use attribute-value pairs to represent instances. An instance is defined by a predetermined group of attributes, such as temperature, and its corresponding value, such as hot. Ideally, we want each attribute to have a finite set of distinct values, like hot, mild, or cold. This makes it easy to construct decision trees. However, more advanced versions of the algorithm can accommodate attributes with continuous numerical values, such as representing temperature with a numerical scale.

**2. The target function has discrete output values:**

The marked objective has distinct outcomes. The decision tree method is ordinarily employed for categorizing Boolean examples, such as yes or no. Decision tree approaches can be readily expanded for acquiring functions with beyond dual conceivable outcome values. A more substantial expansion lets us gain knowledge about aimed objectives with numeric outputs, although the practice of decision trees in this framework is comparatively rare.

**3. Disjunctive descriptions may be required:**

Decision trees naturally represent disjunctive expressions.

**4.The training data may contain errors:**

“Techniques of decision tree learning demonstrate high resilience towards discrepancies, including inconsistencies in categorization of sample cases and discrepancies in the feature details that characterize these cases.”

**5. The training data may contain missing attribute values:**

In certain cases, the input information designed for training might have absent characteristics. Employing decision tree approaches can still be possible despite experiencing unknown features in some training samples. For instance, when considering the level of humidity throughout the day, this information may only be accessible for a specific set of training specimens.

**Practical issues in learning decision trees include:**

* Determining how deeply to grow the decision tree,
* Handling continuous attributes,
* Choosing an appropriate attribute selection measure,
* Handling training data with missing attribute values,
* Handling attributes with differing costs, and
* Improving computational efficiency.

To build the Decision Tree, [CART (Classification and Regression Tree) algorithm](https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/) is used. It works by selecting the best split at each node based on metrics like Gini impurity or information Gain. In order to create a decision tree. Here are the basic steps of the CART algorithm:

1. The root node of the tree is supposed to be the complete training dataset.
2. Determine the impurity of the data based on each feature present in the dataset. Impurity can be measured using metrics like the Gini index or entropy for classification and Mean squared error, Mean Absolute Error, friedman\_mse, or Half Poisson deviance for regression.
3. Then selects the feature that results in the highest information gain or impurity reduction when splitting the data.
4. For each possible value of the selected feature, split the dataset into two subsets (left and right), one where the feature takes on that value, and another where it does not. The split should be designed to create subsets that are as pure as possible with respect to the target variable.
5. Based on the target variable, determine the impurity of each resulting subset.
6. For each subset, repeat steps 2–5 iteratively until a stopping condition is met. For example, the stopping condition could be a maximum tree depth, a minimum number of samples required to make a split or a minimum impurity threshold.
7. Assign the majority class label for classification tasks or the mean value for regression tasks for each terminal node (leaf node) in the tree.

### Classification and Regression Tree algorithm for Classification

Let the data available at node m be Qm and it has nm samples. and tm as the threshold for node m. then, The classification and regression tree algorithm for classification can be written as :

Here,

* H is the measure of impurities of the left and right subsets at node m. it can be entropy or Gini impurity.
* nm is the number of instances in the left and right subsets at node m.

To select the parameter, we can write as:

**Strengths and Weaknesses of the Decision Tree Approach**

The strengths of decision tree methods are:

* Decision trees are able to generate understandable rules.
* Decision trees perform classification without requiring much computation.
* Decision trees are able to handle both continuous and categorical variables.
* Decision trees provide a clear indication of which fields are most important for prediction or classification.
* Ease of use: Decision trees are simple to use and don’t require a lot of technical expertise, making them accessible to a wide range of users.
* Scalability: Decision trees can handle large datasets and can be easily parallelized to improve processing time.
* Missing value tolerance: Decision trees are able to handle missing values in the data, making them a suitable choice for datasets with missing or incomplete data.
* Handling non-linear relationships: Decision trees can handle non-linear relationships between variables, making them a suitable choice for complex datasets.
* Ability to handle imbalanced data: Decision trees can handle imbalanced datasets, where one class is heavily represented compared to the others, by weighting the importance of individual nodes based on the class distribution.

**The weaknesses of decision tree methods :**

* Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
* Decision trees are prone to errors in classification problems with many classes and a relatively small number of training examples.
* Decision trees can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.
* Decision trees are prone to overfitting the training data, particularly when the tree is very deep or complex. This can result in poor performance on new, unseen data.
* Small variations in the training data can result in different decision trees being generated, which can be a problem when trying to compare or reproduce results.
* Many decision tree algorithms do not handle missing data well, and require imputation or deletion of records with missing values.
* The initial splitting criteria used in decision tree algorithms can lead to biased trees, particularly when dealing with unbalanced datasets or rare classes.
* Decision trees are limited in their ability to represent complex relationships between variables, particularly when dealing with nonlinear or interactive effects.
* Decision trees can be sensitive to the scaling of input features, particularly when using distance-based metrics or decision rules that rely on comparisons between values.

# Supervised and Unsupervised learning

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Supervised learning and unsupervised learning are two main types of[machine learning](https://www.geeksforgeeks.org/machine-learning/).

In [supervised learning](https://www.geeksforgeeks.org/supervised-machine-learning/), the machine is trained on a set of labeled data, which means that the input data is paired with the desired output. The machine then learns to predict the output for new input data. Supervised learning is often used for tasks such as classification, regression, and object detection.

In unsupervised learning, the machine is trained on a set of unlabeled data, which means that the input data is not paired with the desired output. The machine then learns to find patterns and relationships in the data. Unsupervised learning is often used for tasks such as [clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/), dimensionality reduction, and anomaly detection.

## **What is Supervised learning?**

Supervised learning is a type of [machine learning algorithm](https://www.geeksforgeeks.org/machine-learning-algorithms/) that learns from labeled data. Labeled data is data that has been tagged with a correct answer or classification.

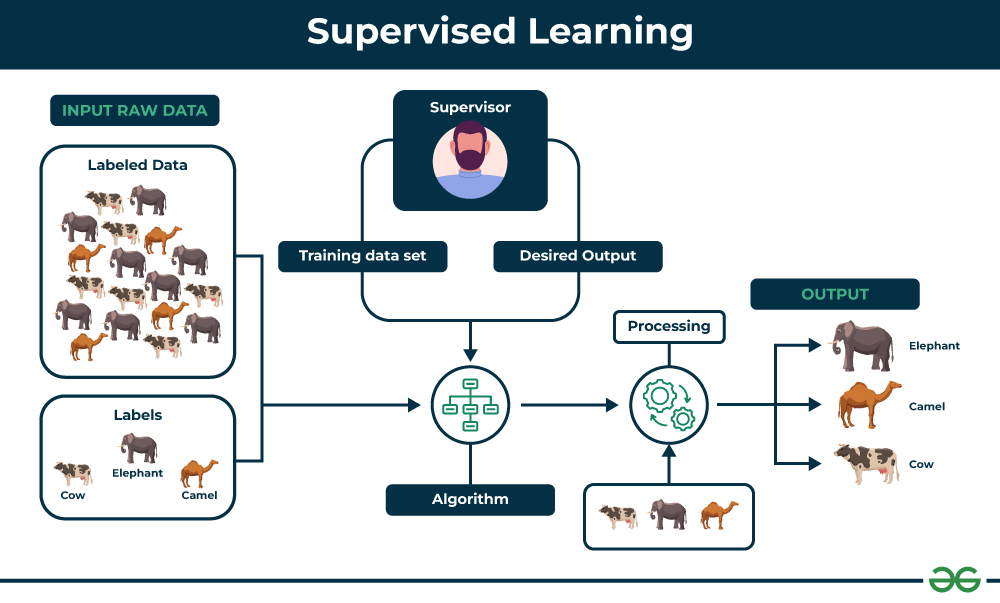
Top VS Code Extensions for Web Developer in 2024!! GeeksforGeeks

**55**

01:10/17:23

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

For example, a labeled dataset of images of Elephant, Camel and Cow would have each image tagged with either “Elephant” , “Camel”or “Cow.”



**Key Points:**

* Supervised learning involves training a machine from labeled data.
* Labeled data consists of examples with the correct answer or classification.
* The machine learns the relationship between inputs (fruit images) and outputs (fruit labels).
* The trained machine can then make predictions on new, unlabeled data.

**Example:**

Let’s say you have a fruit basket that you want to identify. The machine would first analyze the image to extract features such as its shape, color, and texture. Then, it would compare these features to the features of the fruits it has already learned about. If the new image’s features are most similar to those of an apple, the machine would predict that the fruit is an apple.

**For instance**, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

* If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
* If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.

Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

## Types of Supervised Learning

Supervised learning is classified into two categories of algorithms:

* [**Regression**:](https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/) A regression problem is when the output variable is a real value, such as “dollars” or “weight”.
* [**Classification**:](https://www.geeksforgeeks.org/getting-started-with-classification/) A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.

### ****1- Regression****

Regression is a type of supervised learning that is used to predict continuous values, such as house prices, stock prices, or customer churn. Regression algorithms learn a function that maps from the input features to the output value.

Some common [regression algorithms](https://www.geeksforgeeks.org/types-of-regression-techniques/) include:

* Linear Regression
* Polynomial Regression
* Support Vector Machine Regression
* Decision Tree Regression
* Random Forest Regression

### ****2- Classification****

Classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not, or whether a medical image shows a tumor or not. Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes.

Some common[classification algorithms](https://www.geeksforgeeks.org/top-6-machine-learning-algorithms-for-classification/) include:

* Logistic Regression
* Support Vector Machines
* Decision Trees
* Random Forests
* Naive Baye

### Evaluating Supervised Learning Models

Evaluating supervised learning models is an important step in ensuring that the model is accurate and generalizable. There are a number of different [metrics](https://www.geeksforgeeks.org/metrics-for-machine-learning-model/) that can be used to evaluate supervised learning models, but some of the most common ones include:

#### For Regression

* **Mean Squared Error (MSE):** MSE measures the average squared difference between the predicted values and the actual values. Lower MSE values indicate better model performance.
* **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, representing the standard deviation of the prediction errors. Similar to MSE, lower RMSE values indicate better model performance.
* **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and the actual values. It is less sensitive to outliers compared to MSE or RMSE.
* **R-squared (Coefficient of Determination):** R-squared measures the proportion of the variance in the target variable that is explained by the model. Higher R-squared values indicate better model fit.

#### For Classification

* **Accuracy:** Accuracy is the percentage of predictions that the model makes correctly. It is calculated by dividing the number of correct predictions by the total number of predictions.
* **Precision:** Precision is the percentage of positive predictions that the model makes that are actually correct. It is calculated by dividing the number of true positives by the total number of positive predictions.
* **Recall:** Recall is the percentage of all positive examples that the model correctly identifies. It is calculated by dividing the number of true positives by the total number of positive examples.
* **F1 score:** The F1 score is a weighted average of precision and recall. It is calculated by taking the harmonic mean of precision and recall.
* **Confusion matrix:** A confusion matrix is a table that shows the number of predictions for each class, along with the actual class labels. It can be used to visualize the performance of the model and identify areas where the model is struggling.

### ****Applications of Supervised learning****

Supervised learning can be used to solve a wide variety of problems, including:

* **Spam filtering:** Supervised learning algorithms can be trained to identify and classify spam emails based on their content, helping users avoid unwanted messages.
* **Image classification:** Supervised learning can automatically classify images into different categories, such as animals, objects, or scenes, facilitating tasks like image search, content moderation, and image-based product recommendations.
* **Medical diagnosis:** Supervised learning can assist in medical diagnosis by analyzing patient data, such as medical images, test results, and patient history, to identify patterns that suggest specific diseases or conditions.
* **Fraud detection:** Supervised learning models can analyze financial transactions and identify patterns that indicate fraudulent activity, helping financial institutions prevent fraud and protect their customers.
* **Natural language processing (NLP):** Supervised learning plays a crucial role in NLP tasks, including sentiment analysis, machine translation, and text summarization, enabling machines to understand and process human language effectively.

### ****Advantages of Supervised learning****

* Supervised learning allows collecting data and produces data output from previous experiences.
* Helps to optimize performance criteria with the help of experience.
* Supervised machine learning helps to solve various types of real-world computation problems.
* It performs classification and regression tasks.
* It allows estimating or mapping the result to a new sample.
* We have complete control over choosing the number of classes we want in the training data.

### ****Disadvantages of Supervised learning****

* Classifying big data can be challenging.
* Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
* Supervised learning cannot handle all complex tasks in Machine Learning.
* Computation time is vast for supervised learning.
* It requires a labelled data set.
* It requires a training process.

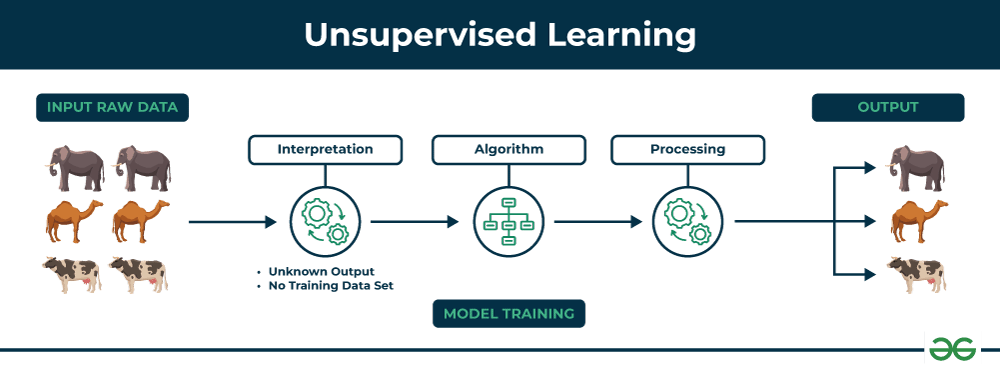
## **What is Unsupervised learning?**

Unsupervised learning is a type of machine learning that learns from unlabeled data. This means that the data does not have any pre-existing labels or categories. The goal of unsupervised learning is to discover patterns and relationships in the data without any explicit guidance.

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

You can use unsupervised learning to examine the animal data that has been gathered and distinguish between several groups according to the traits and actions of the animals. These groupings might correspond to various animal species, providing you to categorize the creatures without depending on labels that already exist.



**Key Points**

* Unsupervised learning allows the model to discover patterns and relationships in unlabeled data.
* Clustering algorithms group similar data points together based on their inherent characteristics.
* Feature extraction captures essential information from the data, enabling the model to make meaningful distinctions.
* Label association assigns categories to the clusters based on the extracted patterns and characteristics.

### Example

Imagine you have a machine learning model trained on a large dataset of unlabeled images, containing both dogs and cats. The model has never seen an image of a dog or cat before, and it has no pre-existing labels or categories for these animals. Your task is to use unsupervised learning to identify the dogs and cats in a new, unseen image.

**For instance**, suppose it is given an image having both dogs and cats which it has never seen.

Thus the machine has no idea about the features of dogs and cats so we can’t categorize it as ‘dogs and cats ‘. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having **dogs** in them and the second part may contain all pics having **cats** in them. Here you didn’t learn anything before, which means no training data or examples.

It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

## Types of Unsupervised Learning

Unsupervised learning is classified into two categories of algorithms:

* **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
* **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

### ****Clustering****

Clustering is a type of unsupervised learning that is used to group similar data points together. [Clustering algorithms](https://www.geeksforgeeks.org/clustering-in-machine-learning/) work by iteratively moving data points closer to their cluster centers and further away from data points in other clusters.

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

**Clustering Types:-**

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis
6. Gaussian Mixture Models (GMMs)
7. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

### ****Association rule learning****

Association rule learning is a type of unsupervised learning that is used to identify patterns in a data. [Association rule](https://www.geeksforgeeks.org/association-rule/)learning algorithms work by finding relationships between different items in a dataset.

Some common association rule learning algorithms include:

* Apriori Algorithm
* Eclat Algorithm
* FP-Growth Algorithm

### Evaluating Non-Supervised Learning Models

Evaluating non-supervised learning models is an important step in ensuring that the model is effective and useful. However, it can be more challenging than evaluating supervised learning models, as there is no ground truth data to compare the model’s predictions to.

There are a number of different metrics that can be used to evaluate non-supervised learning models, but some of the most common ones include:

* **Silhouette score:** The silhouette score measures how well each data point is clustered with its own cluster members and separated from other clusters. It ranges from -1 to 1, with higher scores indicating better clustering.
* **Calinski-Harabasz score:** The Calinski-Harabasz score measures the ratio between the variance between clusters and the variance within clusters. It ranges from 0 to infinity, with higher scores indicating better clustering.
* **Adjusted Rand index:** The adjusted Rand index measures the similarity between two clusterings. It ranges from -1 to 1, with higher scores indicating more similar clusterings.
* **Davies-Bouldin index:** The Davies-Bouldin index measures the average similarity between clusters. It ranges from 0 to infinity, with lower scores indicating better clustering.
* **F1 score:** The F1 score is a weighted average of precision and recall, which are two metrics that are commonly used in supervised learning to evaluate classification models. However, the F1 score can also be used to evaluate non-supervised learning models, such as clustering models.

### Application ****of Unsupervised learning****

Non-supervised learning can be used to solve a wide variety of problems, including:

* Anomaly detection: Unsupervised learning can identify unusual patterns or deviations from normal behavior in data, enabling the detection of fraud, intrusion, or system failures.
* Scientific discovery: Unsupervised learning can uncover hidden relationships and patterns in scientific data, leading to new hypotheses and insights in various scientific fields.
* Recommendation systems: Unsupervised learning can identify patterns and similarities in user behavior and preferences to recommend products, movies, or music that align with their interests.
* Customer segmentation: Unsupervised learning can identify groups of customers with similar characteristics, allowing businesses to target marketing campaigns and improve customer service more effectively.
* Image analysis: Unsupervised learning can group images based on their content, facilitating tasks such as image classification, object detection, and image retrieval.

### Advantages ****of Unsupervised learning****

* It does not require training data to be labeled.
* Dimensionality reduction can be easily accomplished using unsupervised learning.
* Capable of finding previously unknown patterns in data.
* Unsupervised learning can help you gain insights from unlabeled data that you might not have been able to get otherwise.
* Unsupervised learning is good at finding patterns and relationships in data without being told what to look for. This can help you learn new things about your data.

### Disadvantages ****of Unsupervised learning****

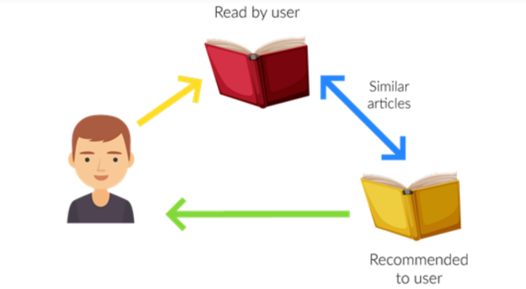
* Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
* The results often have lesser accuracy.
* The user needs to spend time interpreting and label the classes which follow that classification.
* Unsupervised learning can be sensitive to data quality, including missing values, outliers, and noisy data.
* Without labeled data, it can be difficult to evaluate the performance of unsupervised learning models, making it challenging to assess their effectiveness.

## **Supervised vs. Unsupervised Machine Learning**

| **Parameters** | **Supervised machine learning** | **Unsupervised machine learning** |
| --- | --- | --- |
| **Input Data** | Algorithms are trained using labeled data. | Algorithms are used against data that is not labeled |
| **Computational Complexity** | Simpler method | Computationally complex |
| **Accuracy** | Highly accurate | Less accurate |
| **No. of classes** | No. of classes is known | No. of classes is not known |
| **Data Analysis** | Uses offline analysis | Uses real-time analysis of data |
| **Algorithms used** | Linear and Logistics regression, Random forest, multi-class classification, decision tree, Support Vector Machine, Neural Network, etc. | K-Means clustering, Hierarchical clustering, KNN, Apriori algorithm, etc. |
| **Output** | Desired output is given. | Desired output is not given. |
| **Training data** | Use training data to infer model. | No training data is used. |
| **Complex model** | It is not possible to learn larger and more complex models than with supervised learning. | It is possible to learn larger and more complex models with unsupervised learning. |
| **Model** | We can test our model. | We can not test our model. |
| **Called as** | Supervised learning is also called classification. | Unsupervised learning is also called clustering. |
| **Example** | Example: Optical character recognition. | Example: Find a face in an image. |
| **Supervision** | supervised learning needs supervision to train the model. | Unsupervised learning does not need any supervision to train the model. |

# What is Recommendation System?

A recommendation system is a type of machine learning system that provides personalized recommendations to users based on their past behaviors, preferences, and patterns. It is a subclass of information filtering systems that use algorithms to recommend items to users based on their interests or behaviors.



Recommendation systems are widely used in e-commerce, social media, entertainment, and other online platforms to increase user engagement and retention, improve customer satisfaction, and drive sales and revenue.

# How To Work Recommendation System?

here are the four steps of how recommendation systems work:

1. **Collecting user data:** The first step in building a recommendation system is to collect user data. This can include user ratings, reviews, clickstream data, purchase history, and other behavioral data. The data can be collected either explicitly, through user surveys or feedback forms, or implicitly, through user interactions with the platform.
2. **Storing the data:** Once the user data is collected, it needs to be stored in a database or data warehouse for analysis. The data can be stored in a structured or unstructured format, depending on the type and volume of the data.
3. **Analyzing the data:** The next step is to analyze the user data to identify patterns and trends. This can be done using various data analysis techniques like clustering, classification, and regression. The goal is to understand the user’s preferences, behaviors, and interests, and to use this information to make personalized recommendations.
4. **Filtering and recommending:** The final step is to filter the data and make recommendations to the user. This can be done using various recommendation algorithms, such as collaborative, content-based, and hybrid filtering. The algorithm uses the user data and the analysis results to generate a list of recommended items the user will likely be interested in. The recommendations are then presented to the user in a personalized way, such as through a recommendation widget, email, or push notification.

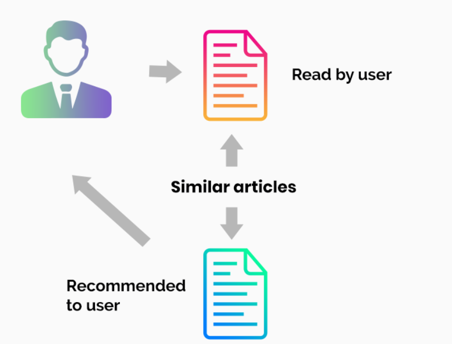
These four steps are the basic components of most recommendation systems, and the specific implementation details may vary depending on the type of system and the application domain.

# Types of Recommendation Systems

There are three main types of recommendation systems

## **→ Content-Based Filtering**

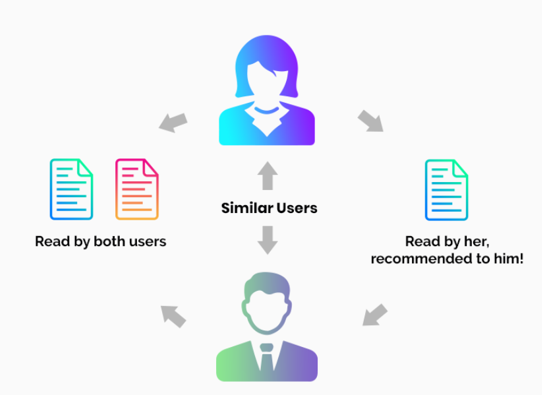
Content-based recommendation systems recommend items to users based on their past preferences and behaviors. This type of system analyzes the user’s historical data, such as their search history, browsing history, or purchase history, and recommends items that are similar to the ones the user has interacted with before.



For example, if a user has watched several action movies in the past, a content-based recommendation system might recommend similar action movies to the user. if a user likes to watch movies such as Iron Man, the recommender system recommends movies of the superhero genre or films describing Tony Stark.

## **→ Collaborative Filtering**

Collaborative filtering recommendation systems recommend items to users based on the preferences and behaviors of other similar users. This type of system analyzes the user’s historical data, as well as the data of other users with similar preferences, and recommends items that similar users have liked or interacted with before. For example, if two users have similar purchase histories, a collaborative filtering recommendation system might recommend items that one user has purchased to the other user.



For example, if user A likes Apples, Bananas, and Mango while user B likes Apples, Bananas, and Jackfruit, they have similar interests. So, it is highly likely that A would like Jackfruit and B would enjoy Mango. This is how collaborative filtering takes place.

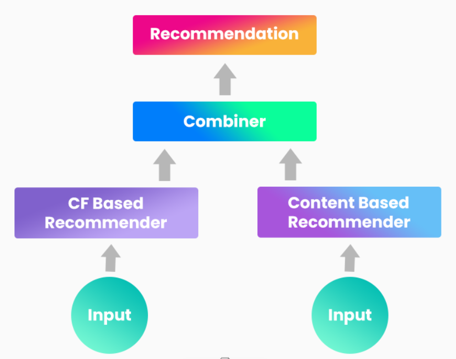
Two kinds of collaborative filtering techniques used are:

* User-User collaborative filtering
* Item-Item collaborative filtering

**User-User collaborative filtering** is a type of recommendation system that makes predictions for a user based on the preferences of similar users. It works by finding users with similar tastes and recommending items they liked to the target user. **Item-Item collaborative filtering,**on the other hand, recommends items to a user based on the preferences for similar items. It works by identifying items that are similar to the ones a user has liked in the past and recommending them to the user.

## **→ Hybrid Recommendation Systems**

Hybrid recommendation systems combine both content-based and collaborative filtering techniques to provide more accurate and diverse recommendations. This type of system uses a combination of user data, item data, and other contextual information to generate recommendations.

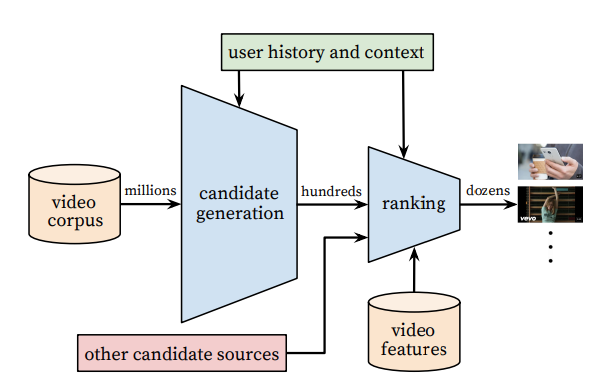
hybrid recommendation system might use content-based filtering to recommend items that are similar to the ones the user has interacted with before, and collaborative filtering to recommend items that other similar users have liked or interacted with. By combining the strengths of both approaches, hybrid recommendation systems can provide more accurate and diverse recommendations than either content-based or collaborative filtering alone.

Netflix is an excellent case in point for a hybrid recommendation system. It makes recommendations by juxtaposing users’ watching and searching habits and finding similar users on that platform. This way, Netflix uses collaborative filtering.

By recommending such shows/movies that share similar traits with those rated highly by the user, Netflix uses content-based filtering. They can also veto the common issues in recommendation systems, such as cold start and data insufficiency issues.

# How does the YouTube algorithm work?

The YouTube recommendation algorithm is a complex system that uses a combination of collaborative filtering, deep learning, and other techniques to personalize video recommendations for each user.



Here are some key factors that the algorithm takes into account:

1. **User engagement:** The algorithm considers the videos a user has watched, liked, commented on, or shared to understand their preferences and interests.
2. **Similarity:** The algorithm identifies videos that are similar to the user’s viewing history, such as videos from the same channel or related topics.
3. **Popularity:** The algorithm takes into account the overall popularity of a video, such as the number of views, likes, and comments.
4. **Freshness:** The algorithm also considers the recency of the video to ensure that users are recommended the latest and most relevant content.
5. **Diversity:** The algorithm tries to recommend a diverse range of content to ensure that users are exposed to new and interesting videos outside of their typical viewing habits.

Overall, the YouTube algorithm is designed to provide personalized and engaging recommendations to each user while keeping them engaged and active on the platform.

## Issues in Recommendation Systems:

## 1Ethical issues

One of the ethical issues of recommender systems is the potential for bias and discrimination. Bias can arise from the data, the algorithms, or the users themselves, and can lead to unfair or inaccurate recommendations that favor or exclude certain groups, opinions, or values. For example, a recommender system that relies on historical data may perpetuate existing stereotypes or inequalities, or a recommender system that uses collaborative filtering may create echo chambers or filter bubbles that limit the diversity and exposure of users to alternative perspectives or information. Moreover, some recommender systems may manipulate or influence the users' choices or behavior for commercial or political purposes, without their consent or awareness.

## 2Social issues

Another issue of recommender systems is the social impact they have on the users and the society. Recommender systems can affect the users' autonomy, privacy, trust, and well-being, depending on how they are designed, implemented, and used. For instance, a recommender system that collects and analyzes the users' personal data may violate their privacy or expose them to security risks, such as data breaches or identity theft. A recommender system that provides inaccurate or misleading recommendations may erode the users' trust or confidence in the system or the provider. A recommender system that induces excessive or addictive consumption or engagement may harm the users' mental or physical health or social relationships.

## 3Technical issues

A third issue of recommender systems is the technical challenges they face in developing and maintaining high-quality and robust systems. Recommender systems need to deal with various problems, such as data sparsity, scalability, cold start, diversity, serendipity, explainability, and evaluation. For example, a recommender system that has insufficient or sparse data may not be able to provide relevant or personalized recommendations, or a recommender system that has to handle a large number of users or items may not be able to provide fast or efficient recommendations. Furthermore, recommender systems need to balance trade-offs between different objectives or criteria, such as accuracy, diversity, novelty, transparency, and user satisfaction.

## 4Legal issues

A fourth issue of recommender systems is the legal implications they have in different jurisdictions and contexts. Recommender systems may be subject to various laws and regulations that govern the collection, processing, storage, and sharing of personal data, such as the General Data Protection Regulation (GDPR) in the European Union, or the California Consumer Privacy Act (CCPA) in the United States. They may also be liable for the consequences or damages that result from their recommendations, such as product liability, consumer protection, copyright infringement, or defamation. Therefore, recommender systems need to comply with the applicable laws and ethical principles, and ensure that they respect the rights and interests of the users and other stakeholders.

## 5Here’s what else to consider

This is a space to share examples, stories, or insights that don’t fit into any of the previous sections. What else would you like to add?